Auto-adaptive averaging: Detecting artifacts in event-related potential data using a fully automated procedure

Talsma, D.

published in
Psychophysiology
2008

DOI (link to publisher)
10.1111/j.1469-8986.2007.00612.x

document version
Publisher's PDF, also known as Version of record

Psychophysiology, 45 (2008), 216–228. Blackwell Publishing Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2007.00612.x

Auto-adaptive averaging: Detecting artifacts in event-related potential data using a fully automated procedure

DURK TALSMA
Cognitive Psychology Department, Vrije Universiteit, Amsterdam, the Netherlands

Abstract

The auto-adaptive averaging procedure proposed here classifies artifacts in event-related potential data by optimizing the signal-to-noise ratio. This method rank orders single trials according to the impact of each trial on the ERP average. Then, the minimum residual background noise level in the ERP data is determined at each step in the averaging process. Trials having a negative impact on the residual background noise are discarded from the averaging procedure. Simulations showed that ERP estimates obtained by the auto-adaptive averaging procedure were either better or comparable to those obtained by single trial artifact detection methods at their most optimum configuration, in particular during long duration artifacts. Experimental data from a working memory task further illustrate the effectiveness of the method.

Descriptors: Artifacts, Auto-adaptive averaging, Electroencephalography, Event-related potentials

Event-related potentials (ERPs) are transient electrical voltage fluctuations that the brain elicits in response to stimuli, behavioral responses, or a wide range of cognitive processes. ERPs are embedded in electroencephalographic (EEG) recordings, and are typically analyzed by averaging together segments of EEG activity (i.e., trials) that are each time-locked to the eliciting event. The underlying assumption of the averaging procedure is that the EEG signal at each trial is composed of a constant part reflecting the ERP activity and a randomly fluctuating part, constituting the concerted background activity of ongoing brain processes that are not specifically related to the event of interest. Because the randomly fluctuating background activity is by definition not correlated with the processing of the event of interest, it would theoretically average out completely should an infinite number of trials be available, leaving only the constant ERP in the remaining average. Ideally, if the ERP signal is constant, the signal-to-noise ratio of the average improves by the square root of the number of trials included in the average (beim Graben, 2001; Niedermeyer & Lopez-De Silva, 1993; see also the Appendix).

ERPs are composed of a series of components, which are described by their polarity in combination with either a rank order or an estimate of the average latency of the component. For instance, P1 would indicate the first positive component and N170 would refer to a negative polarity component at 170 ms after the eliciting event. Early-latency (<200 ms after stimulus onset) components are typically related to the initial processing of sensory events and are characterized by a low amplitude (of about 1–5 μV) and a relatively high frequency. Longer latency (i.e., 200 ms and later) components are typically of a higher amplitude (10–20 μV), characterized by a low frequency, and related to higher stages of cognitive processing. In all cases, these amplitudes are significantly lower than the amplitude of the average background EEG, which is typically on the order of about 50 μV. Consequently, a large number of trials are required to reduce the signal-to-noise ratio of the ERP to acceptable levels. The precise number of trials required varies somewhat, depending on the component of interest. In general, the smaller the amplitude of a component, the larger the required number of trials is. For instance, the estimation of a large amplitude component such as the P3, or the negative slow wave, can be accomplished using about 40–80 trials (Bosch, Mecklinger, & Friederici, 2001; Johnson, 1989; Pelosi & Blumhardt, 1999; Ruchkin et al., 1997; Scheffers, Johnson, & Ruchkin, 1991), whereas a reliable estimation of the early sensory components typically employs a significantly higher number of trials per ERP average (Hickey, McDonald, & Theeuwes, 2006; Molholm, Ritter, Javitt, & Foxe, 2002; Talsma, Doty, & Woldorff, 2007; Talsma, Kok, & Ridderinkhof, 2006; Talsma & Woldorff, 2005b; Yago, Escera, Alho, & Giard, 2001).

Source code of the implemented version of the auto-adaptive averaging method can be downloaded from the CVS repository at the sourceforge Web site (http://sourceforge.net/projects/erp). I thank my colleagues at our department for the helpful discussions leading up to the development of this method. I further thank Sofia Diamantoupoulou for collecting the working memory data and two anonymous reviewers for their helpful comments on an earlier draft of this article.

Address reprint requests to: Durk Talsma, Cognitive Psychology Department, Vrije Universiteit, Van den Boechorststraat 1, 1081 BT Amsterdam, The Netherlands. E-mail: d.talsma@psy.vu.nl

Requiring such large numbers of trials in each average has the practical consequence that ERP recording sessions are typically rather long. Experimental design requirements can place a significant restriction on the number of trials that can practically be included in the experiment. During off-line analysis, the actual number of trials available for inclusion in ERP averages is even further reduced due to the exclusion of trials that are contaminated by recording artifacts. In earlier work, Marty Woldorff and I defined artifacts as "occurrences of electrical activity that can be recorded by EEG equipment, which is not originating from cerebral sources and is either clearly distinguishable from the recorded background EEG or substantially large enough to modify the observed ERP waveform from its true waveform" (Talsma & Woldorff, 2005a). Following this definition, those occurrences of electrical activity that are large enough to modify the observed ERP waveform should be eliminated from the data as well as possible.

In the remainder of this article, I make a distinction between ocular artifacts and instrumentation artifacts, with a particular emphasis on instrumentation artifacts. Due to the fact that eye movements and blinks are one among the major sources of artifacts in ERP data, ocular activity is typically recorded on dedicated EOG channels, which can be used to quantify ocular activity. Due to its known spatio-temporal characteristics, the influence of ocular artifacts is well understood, and several methods exist for the correction of eyeblink and eye-movement artifacts in ERPs (Berg & Scherg, 1994; Croft & Barry, 1998, 2000, 2002; Gratton, Coles, & Donchin, 1983; Jung et al., 2000; van den Berg-Lenssen, Brunia, & Blom, 1989; Woestenburg, Verbaten, & Slangen, 1983).

In contrast, instrumentation artifacts can originate from a large number of sources, including, but not limited to, muscle activity, movement, inadequate shielding, and equipment failures. These artifacts can be classified on the basis of their origin and the type of distortion that they exert on the ERP signal. Spike artifacts are characterized by a more or less transient fluctuation of a relatively high voltage. This type of artifact can be caused by muscle activity, body movements during the recording session, a sudden change in electrode contact, or interference from other electrical sources such as electrocardiographic activity or pulsating arteries (Fisch, 1991; Talsma & Woldorff, 2005a).

Slow drift potentials, such as those generated by skin potentials or incorrectly placed electrodes can be particularly problematic, because their frequency content is of the same order as that of the slow-wave brain activity that is related to anticipatory processes (CNV; Walter, Cooper, Aldridge, McCallum, & Winter, 1964), direction of attention (Hopf & Mangun, 2000), or working memory (Drew, McCollough, & Vogel, 2006; Klaver, Talsma, Wijers, Heinze, & Mulder, 1999). For this particular reason, some research groups revert to installing custom-built filters in commercially available hardware (Grent-'t-Jong & Woldorff, 2007), or use custom-built amplifiers (Klaver et al., 1999; Talsma, Wijers, Klaver, & Mulder, 2001), having high-pass frequency cutoffs as low as 0.01 Hz, in order to filter out the drifts but leave the slow wave intact. When using such low high-pass settings is not possible, recordings can be made in DC mode, that is, without the application of a high-pass filter during data acquisition. A significant drawback in DC mode recording, however, is that the recorded EEG signals are extremely sensitive to slow drifts due to polarizing electrodes or unstable connections, resulting in "dead" (clipping) signals and unpredictable interactions with the digitization hardware.

In practice, many artifacts can be found that share characteristics between these major two classes of instrumentation artifacts. For instance, electrodes that are beginning to lose contact can show jumps, without returning immediately to baseline levels, as a result of a sudden change in impedance. Likewise, these electrodes can show erratic random behavior that cannot be fully classified using just amplitude changes or linear drift components.

A major goal of the ERP signal-averaging procedure is to maximize the signal-to-noise ratio by including as many trials as possible. For this reason, one should aim to achieve an optimum balance between maintaining a high number of trials and removing artifacts. Excluding a few trials with large artifacts will help improve the overall quality of the data, whereas excluding many trials containing tiny artifacts will not. In the latter case the signal-to-noise ratio of the ERPs will remain poor, due to the low number of trials that the ERP average is composed of (Talsma & Woldorff, 2005a).

Many commercially available automated artifact classification algorithms work on the basis of determining the peak amplitude in EEG signals. For instance, artifact detection procedures are oftentimes described by saying that EEG activity exceeding some threshold from the mean was considered to be artifactual. Although this is generally true for spike artifacts, the use of such a cutoff procedure can be problematic in case of slow drifting potentials. In the latter case, although average EEG activity may indeed exceed the cutoff threshold, due to a preceding drift, the particular segment of EEG activity by itself may be free of artifacts. It is possible to detect drift potentials using linear regression (Hennighausen, Heil, & Rosler, 1993; Talsma & Woldorff, 2005a); however, it should be noted that these methods must be applied with care, as using incorrect cutoff criteria can still result in the selective inclusion of small slow drift potentials in one direction, which are no longer countered by larger drift trials of opposite polarity. Oftentimes this may result in a "cleaned-up" signal that has a net drift that is substantially larger than that of the uncorrected signal (see Talsma & Woldorff, 2005a, for an example of this effect).

As illustrated, the existing automated methods work reasonably well for the identification of spikelike artifacts, which are typically magnitudes larger than the background EEG signal. However, these methods do not work so well in identifying other types of artifacts, such as slow drifting potentials or noisy electrodes. In addition, a common disadvantage of these methods is that they require extensive user configuration, and oftentimes seemingly arbitrary decisions regarding cutoff criteria.

The main aim of the present article is to introduce a new method that takes a fundamentally different approach. This method, termed "auto-adaptive averaging," finds artifactual trials by determining the impact of each single trial on the ERP average, rank orders trials according to impact, and estimates the minimum attainable residual noise term for each ERP average. The auto-adaptive averaging method was tested using simulation studies, and due to the automatic minimum residual noise determination procedure, I expected that the accuracy of the auto-adaptive averaging procedure in determining artifactual trials would be comparable to or exceed that of the established single-trial artifact detection procedures in their most optimal configuration. In addition, I expected that the auto-adaptive averaging algorithm would be able to accomplish this without any user configuration.

Methods

Auto-adaptive Averaging

The auto-adaptive averaging procedure described here is based on the principle that ERPs that are composed of artifact-free trials will be more or less similar independent of which subset of trials are included in the average. The inclusion of artifactual trials, on the other hand, will have a distorting impact on the observed ERP signal. According to this principle, the exclusion of a single artifact-free trial from the full set of trials should have a negligible impact on the observed ERP waveform. In contrast, the exclusion of a trial that does happen to contain an artifact should have an observable impact. Therefore, artifactual trials can be found by computing the impact that each trial has on the observed ERP waveform. In the auto-adaptive averaging procedure, this is obtained by rank ordering trials according to impact on the average followed by determining at which point this impact becomes artifactual. Figure 1 outlines the main components of the auto-adaptive averaging procedure described here.

Theoretical background. The signal-to-noise ratio of an ERP consisting of N trials is expected to improve by a $\sqrt{N}$ compared to that of a single trial. Because the ERP is considered to be time invariant, this improvement in signal-to-noise ratio can be equated to an inverse (i.e. $1/\sqrt{N}$) reduction in noise power. When individual trials differ in background noise, however, due to artifacts or other causes, the estimated noise power of each individual trial has to be considered. Such an estimate can be obtained by averaging across the noise power estimates obtained for each single trial (see the Appendix for details). Figure 2a shows that the inclusion of a relatively noisy trial has an adverse effect on the estimated noise level in the averaged signal. Although this inverse impact is particularly large when the average still consists of a low number of trials, Figure 2a shows that the estimated noise levels continue to decrease as more and more trials are included in the average, even including artifactual ones, but that the final noise power estimates in artifact-containing ERPs remains higher than that of artifact-free data.

As shown in Figure 2a, averaging all the trials still results in a gradually decreasing noise estimate, with intermittent artifact-related increases of the noise contribution estimates. This observation could therefore be used to falsely infer that the optimum ERP average could be obtained by including all trials in the average, including the artifactual ones. According to Figure 2a, this would lead to the minimally attainable noise power and therefore to the highest attainable signal-to-noise ratio. Notice, however, that the estimated noise power at this stage is still considerably higher than the estimated noise power that would have resulted in the case of clean EEG epochs (Figure 2a, thin solid line).

Figure 2b shows the same noise-power estimates as those depicted in Figure 2a, but now after rank ordering trials according to the estimated noise power of each individual trial. When averages are constructed by adding trials in the order of increasing noise power, the residual noise power estimates initially decrease, as shown by the deceasing noise contribution plots (Figure 2b, thick solid line). As more and more increasingly noisy trials are added, however, the noise power estimate of the average first reaches a minimum, after which it increases again. Ideally, ERP averaging should terminate at the moment the minimally attainable noise power estimate has been reached.

Impact determination. Although an estimate of the noise power of each trial could be obtained according to the equations given in the Appendix, the rank ordering of trials can in practice be accomplished using a computationally less intensive impact determination procedure. This method is based on obtaining the difference between two averages as the impact measure. The first of these averages consists of the full average, consisting of all trials N. The second average consists of $N-1$ trials. In other words, for each trial i ($1 \leq i \leq N$) of the ensemble, a difference wave is created between the average consisting of all N trials and an average consisting of all trials except trial i. The impact of trial i can then be determined by computing the power contained within the difference wave (averaged across channels), which yields a measure of the overall similarity of the N and $N-i$ averages. This measure is sensitive to relatively long-lasting artifacts, such as drifts and electrode jumps.

Classifying artifactual trials. Having established the impact of each single trial and rank ordered the trials according to impact, the next step is to determine the point at which the relative impact of a single trial should be classified as artifactual. In practice, this is accomplished using two additional averaging steps. First, for each trial i ($1 \leq i \leq N$) of the rank-ordered ensemble, two averages are created. The first of these averages is composed of the rank-ordered trials 1 to i ($\mathrm{Av}_{(i)}$) and the second of the rank-ordered trials 1 to $i-1$ ($\mathrm{Av}_{(i-1)}$). Then, a difference wave is created by subtracting the ($\mathrm{Av}_{(i-1)}$) average from the ($\mathrm{Av}_{(i)}$) average. Finally, the power of this difference wave is taken as an estimate of the power of the residual noise $RN_{(i)}$ in the ERP waveform after averaging trials 1 to i. These residual power estimates can be plotted as a function of the number of rank-ordered trials in the average, yielding residual noise power curves similar to those of Figure 2b. The minimum of this function $RN_{(min)}$ can be taken as the point to determine the point where averaging should be terminated. In the second averaging step, all the trials 1 to $i_{(RN(min))}$ are used to create the final ERP.

Simulations

A total of four different sets of simulations were run. The first of these determined the effectiveness of discarding slow drift trials, the second simulation determined the effectiveness of discarding signal jumps, the third simulation determined the effectiveness of discarding transient spikes, and the fourth simulation determined the effectiveness of discarding random artifacts. EEG, ERPs, and artifacts were simulated using the Electrophysiological Analysis System (EASY) software toolkit developed by the author (released as open source software and available at http://www.sf.net/projects/erp). Thirty-two channels of virtual EEG data were simulated at a 250-Hz sampling frequency, consisting of two layers of Perlin noise (Perlin, 2002), scaled to $\pm$ 50 μV (layer 1) and $\pm$ 10 μV (layer 2).[1] Perlin noise was chosen due to its $1/F$ noise frequency characteristic common in many natural systems, including the EEG (see Figure 3 for examples of simulated EEG with artifacts), and its coherence across multiple spatiotemporal dimensions. Finally, high-frequency random interference noise of $\pm$ 1 μV was inserted by randomly displacing samples. This simulated EEG was scaled between $\pm$ 50 μV. Then, in each block of simulated EEG, a total of 100 simulated

[1] Although the simulation data are computer generated, the common electro-physiological units of microvolts and milliseconds are used throughout the entire article.

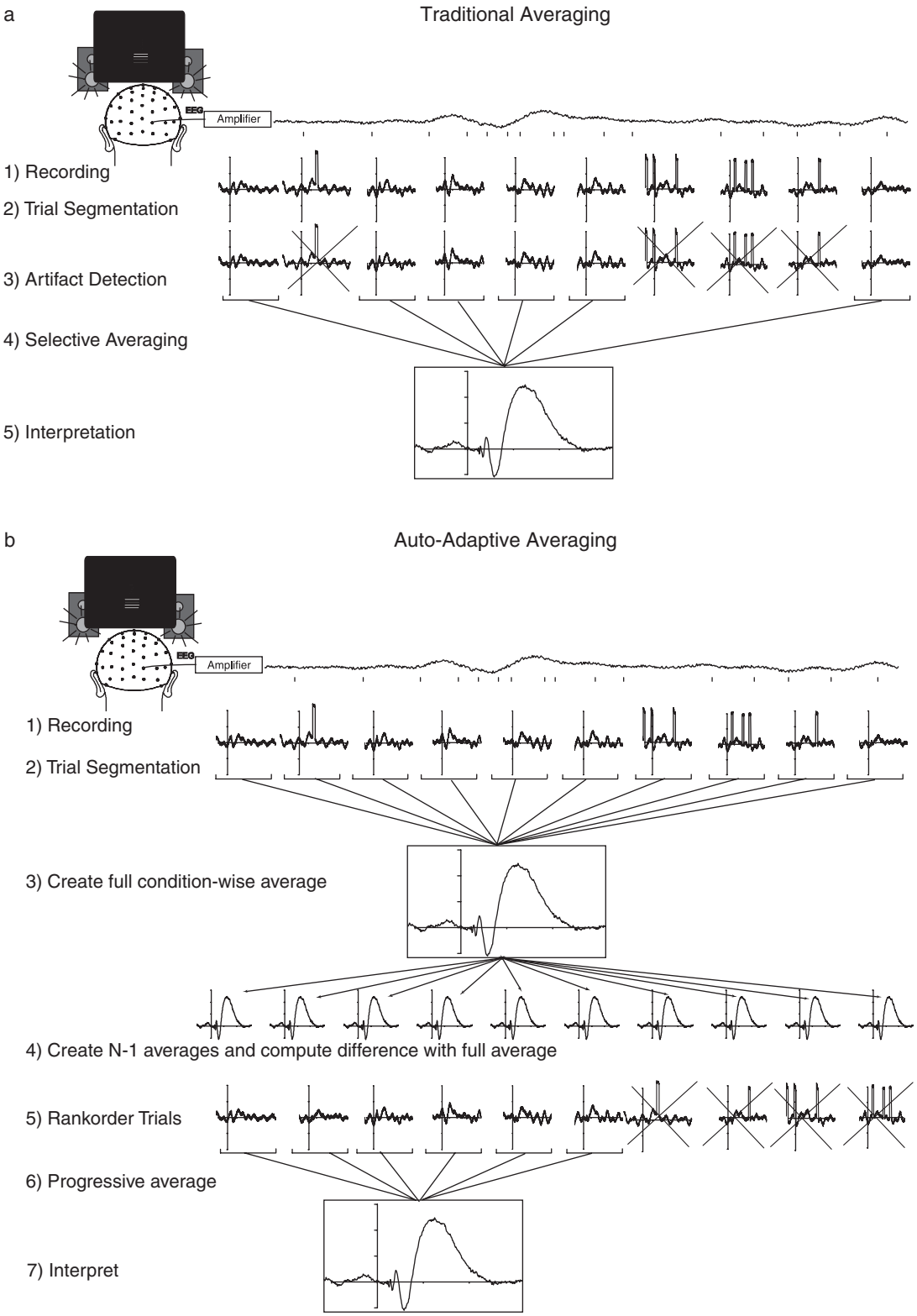

Figure 1. Comparison between traditional averaging procedures and the auto-adaptive averaging procedure. Whereas traditional averaging methods (a) detect artifacts prior to averaging, the artifact detection procedure forms an integral part of the auto-adaptive averaging process (b). After recording and trial segmentation (steps 1 and 2), a full average is created based on all trials, including the artifactual ones (step 3). This full average is subtracted from a second average, where one trial is left out (step 4). The power of this difference wave is taken as a measure of the impact of the omitted trial. Then, trials are rank ordered according to increasing impact (step 5). Finally, the increasingly impacting trials are averaged up to the point where the estimated residual noise in the ERP is reaching a minimum (step 6). In practice, this goal is attained using two averaging passes, one in which the minimum residual noise term is estimated and second pass one in which the final average is assembled. See main text for further details.

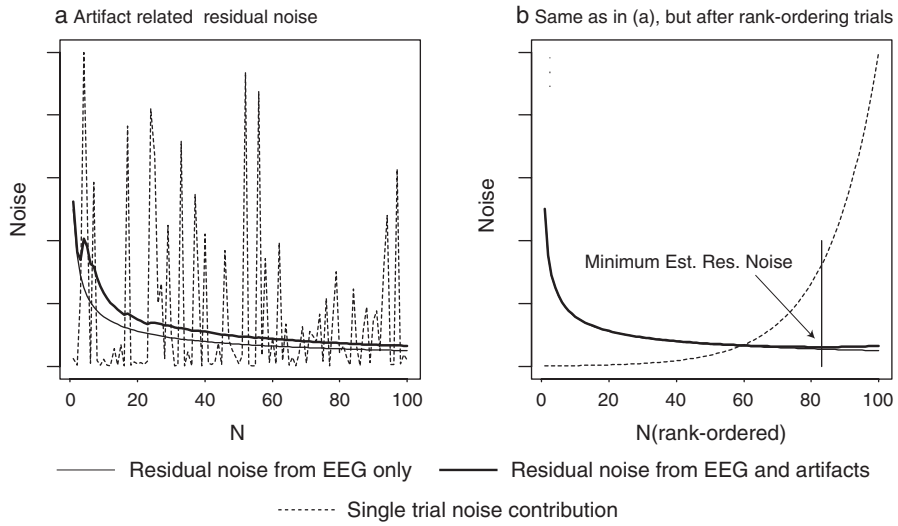

Figure 2. Residual noise plots. a: Estimate of residual noise as a function of the number of trials averaged when artifacts are causing considerable differences in the observed single-trial noise. The inclusion of noisy trials (illustrated by spikes in the dashed line) can cause a temporary increase in residual noise as such trials are included in the average. b: The minimally attainable residual noise can be estimated by rank ordering trials according to impact and including relatively low-impact trials first in the average. As trials with increasing noise are included in the average, residual noise estimates will reach a minimum and then increase again. In both plots, the *x*-axis corresponds to the number of trials used for the residual noise curves (solid lines) as well as to the trial index for the single trial noise estimate $PN(i)$ (dashed line).

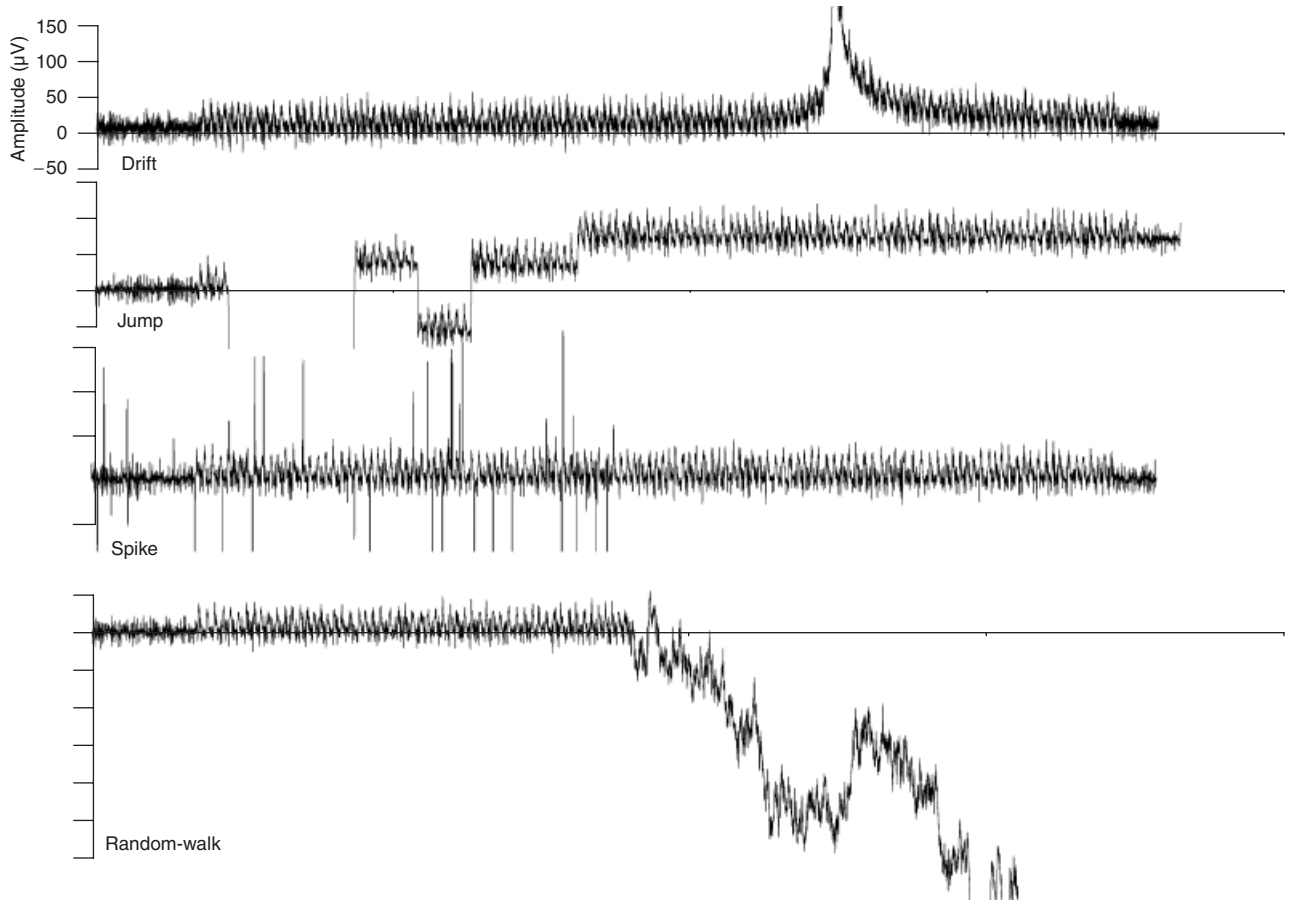

Figure 3. Examples of Perlin-noise based EEG data simulations with drift, jump, spike, and random walk artifacts.

ERPs were inserted at random intervals of 1000 to 1500 ms. These simulated ERPs consisted of a frequency and amplitude modulated sine wave function, with a duration of 256 samples, which was constructed using the following equation:

$$Y[i] = \frac{\left(\sin(256/i)(0.1i^2)(256-i)^2\right)}{1000000}$$

where $0 < i < 257$. This function was chosen as it resembles the characteristics of a typical visual ERP in that it consists of relatively low-amplitude and high-frequency waves at the beginning of the signal and relatively low frequency and high amplitudes (of about 25 μV) at the end. This function is plotted as a reference in each graph in Figures 4–7 (thin dashed line). It should be noted that identical simulated ERPs were inserted at each of the 32 simulated channels.

After the creation of the background noise and the insertion of the simulated ERPs, random artifacts were inserted. Depending on which simulation was run, these artifacts consisted of simulated spikes, channel jumps, drifts, or a random walk displacement, as illustrated in Figure 3.

For each artifact type, 20 blocks of simulated EEG data were created, and for each block 10 different ERP averages were created. The first of these averages was constructed using the auto-adaptive averaging procedure, the second average was created without any artifact-detection methods, and the remaining eight averages were created using single-trials artifact detection methods using increasingly stricter cutoff criteria. For each simulation, the single-trial artifact detection procedures were configured such that the most liberal tests rejected hardly any trials, whereas the most conservative test rejected the majority of trials. Each of these 10 averages was compared to the true simulated ERP, and the power of the residual noise (i.e., the difference between the true and the estimated ERP simulation) was taken as a measure of the quality of the averaging procedure. Lower power values of this difference wave denoted a better estimate of the true simulated ERP. Finally, the 20 power estimates obtained for each type of average (across the 20 blocks of simulated EEG data) were used as a dependent measure in a pairwise *t* test, in which the residual power estimate of the auto-adaptive averaging procedure was compared to one of the other estimates (i.e., single-trial artifact detection method or no artifact detection at all).

Finally, to assess the effectiveness of the method in a multichannel environment, each of these simulations was repeated five times. In consecutive simulations, the number of channels containing artifacts was doubled, resulting in sessions with 2, 4, 8, 16, and 32 artifactual channels. In all simulations, power estimates were obtained across only those channels that had simulated artifacts in them.

Slow drifts. Slow drifts were constructed by creating an initial +300-μV amplitude jump that started at a random time point. Starting from this jump, the signal was simulated to drift back to 0 mean amplitude, by attenuating the initial jump amplitude according to

$$j_{[i]} = j_{[i-1]} - \left(\frac{j^2_{[i-1]}}{j^2_{[0]}}\right), \quad 1 < ; j < ; N, \quad \text{and}$$
$$Y_{[i]} = Y_{[i]} + j_{[i]}$$

where $j_{[0]}$ equals the initial+300-μV amplitude jump and $j_{[i]}$ the attenuated drift at time point *i*. The same procedure was repeated in the reverse direction, starting from the initial jump point, but using a four-times stronger attenuation factor to simulate a fast ramping up followed by a slower return drift. Single-trial artifact detection was accomplished by computing a linear regression line through each epoch and at each channel using a 2-s period. For each average, threshold values were set at ± 5 μV/s, ± 10 μV/s, ± 15 μV/s, ± 20 μV/s, ± 25 μV/s, ± 30 μV/s, ± 35 μV/s, and ± 40 μV/s, respectively.

Signal jumps. For each affected channel, 25 signal jumps were created by offsetting the signal by a random value between ± 200 μV, using a random ramp of 4 to 50 ms. Signal jumps were restricted to occur only in the first half of the simulated EEG signal. Single-trial artifact detection was accomplished by testing for signal deviations across a moving window of 40 ms (Talsma & Woldorff, 2005a), exceeding ± 10 μV, ± 30 μV, ± 50 μV, ± 70 μV, ± 90 μV, ± 110 μV, ± 130 μV, and ± 150 μV, respectively, at each test.

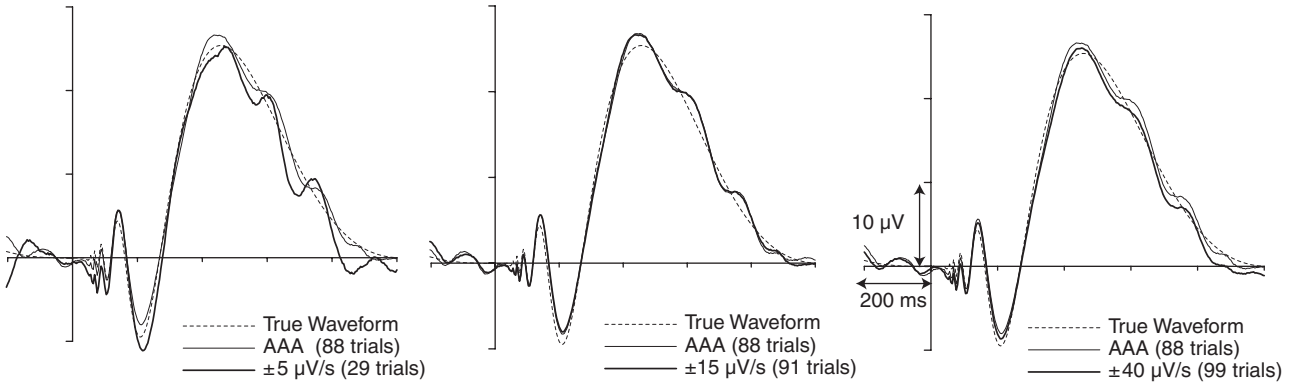

Figure 4. Drift detection simulations. Shown here is a comparison of the auto-adaptive averaging procedure with three single-trial regression-based drift detection results. All results were obtained from the 32-channel condition. For comparison, the true simulated ERP waveform is plotted in each figure. Whereas the single-trial cutoff of ± 5 μV/s was too strict, the ± 40 μV/s cutoff criterion was unable to exclude some drift trials. Performance of the auto-adaptive averaging procedure was comparable to that obtained by that of the ± 15 μV/s single-trial drift detection technique. In all other simulations, the estimates obtained by the auto-adaptive average technique were either equivalent or significantly closer to the true waveform than those obtained by the other averages. AAA: auto-adaptive averaging.

Spikes. For each affected channel, 50 spikes were generated of a random amplitude ranging between ± 150 μV, with a random duration of 50 to 250 ms. Spikes were restricted to occur only in the first half of the simulated EEG signal. Single-trial artifact detection was accomplished by testing for signal deviations across a moving window of 4 ms (equaling two consecutive samples), exceeding ± 10 μV, ± 30 μV, ± 50 μV, ± 70 μV, ± 90 μV, ± 110 μV, ± 130 μV, and ± 150 μV, respectively, at each test.

Random walk artifacts. A random walk simulation was used due to its resemblance to an electrode that is gradually breaking contact. On each affected channel, a random walk artifact was introduced to the second half of the simulated EEG signal by adding a stochastic parameter to the simulated signal that was initially set to zero. At each next sample this parameter could randomly increase or decrease by 2 μV. Single-trial artifact detection was accomplished by testing for signal deviations across a moving window of 100 ms, exceeding ± 10 μV, ± 30 μV, ± 50 μV, ± 70 μV, ± 90 μV, ± 110 μV, ± 130 μV, and ± 150 μV, respectively, at each test.

Experimental Data

EEG recordings were obtained from 8 participants who completed a visual working memory task. ERPs were computed as participants were encoding one, two, three, or four abstract shapes into memory during a 1500-ms memorization interval, followed by a 1500-ms retention interval, and were time-locked to presentation of the memory items. Memory items were presented to both the left visual and right visual field, but only the items in one of the hemifields were memorized, which was indicated by a cue that occurred 1000 ms prior to memory display onset. The data shown in the present article are from a right hemisphere recording site taken while participants were memorizing the left hemifield items (i.e., contralateral to the memorized visual field). EEGs were recorded using a Neuroscan SynAmps amplifier, running in DC mode, using a sample frequency of 500 Hz and a low-pass filter at 100 Hz. No off-line filtering was applied to these data, except for a nine-point moving average to attenuate 50-Hz line interference. A relatively large number of recording artifacts occurred during some sessions, due to polarizing electrodes, which yielded suboptimal results and extensive user configuration of the artifact detection criteria using traditional averaging methods.

Results

Simulations

Slow drifts. Figure 4 compares the averages obtained by the auto-adaptive averaging method to three representative results obtained using single-trial artifact detection methods. Each average shown in Figure 4 was obtained from one block of data containing 100 trials from the simulation run in which all 32 channels were affected by artifacts. When liberal criteria were used in the single-trial artifact detection methods, the overall average was representative of the true simulated ERP, although a small drift component was still visible. Performance statistics across the 10 different artifact detection procedures and five channel sets are given in Table 1. Each table entry represents a Student's *t* statistic that was obtained by comparing the auto-adaptive averaging procedure with one of the single-trial artifact detection procedures. Significant negative *t* values indicate that the auto-adaptive averaging procedure estimated the true ERP waveform with a higher accuracy than that of the corresponding single-trial estimate. As shown by Table 1, performance of the auto-adaptive averaging method was either comparable to the single-trial methods or more accurate.

Channel jumps. Results from the channel jump simulations are shown in Figure 5. As shown in this figure, reasonably adequate estimates of the true simulated ERP could be obtained using both the auto-adaptive averaging method and the single-trial detection methods. Similar to the drift analyses described above, Table 2 summarizes the Student's *t* statistic for the comparison between residual power estimates of the simulated ERPs obtained by the auto-adaptive averaging method and the residual power estimates of the other averages. Close inspection of Table 2 reveals an interesting pattern of results: In general, the auto-adaptive averaging procedure was more accurate in estimating the true waveform compared to the single trial methods that were either too conservative (i.e., ± 10 μV and ± 30 μV) or too liberal (i.e., ± 130 μV, ± 150 μV, or using no artifact detection at all). However, through a small band of cutoff values (± 50 μV to ± 90 μV) artifact detection was actually more accurate using the single-trial amplitude jump method.

Table 1. *Comparison of AAA with Single-Trial Artifact Detections in Drift Simulations*

	Number of artifactual channels				
Type	2	4	8	16	32
None	− 2.80	n.s.	− 2.6	− 2.78	− 1.78
± 5 μV/s	− 8.52	− 8.99	− 12.5	− 11.54	− 6.67
± 10 μV/s	− 2.88	− 3.19	− 4.63	− 4.71	− 4.70
± 15 μV/s	n.s.	n.s.	n.s.	− 2.18	n.s.
± 20 μV/s	− 2.11	n.s.	n.s.	n.s.	n.s.
± 25 μV/s	− 2.18	n.s.	n.s.	− 2.56	n.s.
± 30 μV/s	n.s.	n.s.	n.s.	− 2.92	n.s.
± 35 μV/s	− 2.10	n.s.	n.s.	− 2.88	n.s.
± 40 μV/s	− 2.13	n.s.	n.s.	− 2.80	n.s.

Note. All values are *t* statistics (19 *df*, alpha = .05), comparing the auto-adaptive averaging performance to the single-trial artifact detection methods using the specified cutoff values and number of artifactual channels. Negative *t* values indicate that the auto-adaptive averaging methods estimated the true signal with higher accuracy than the comparison method. AAA: auto-adaptive averaging; None: average obtained without artifact detection, n.s.: not significant.

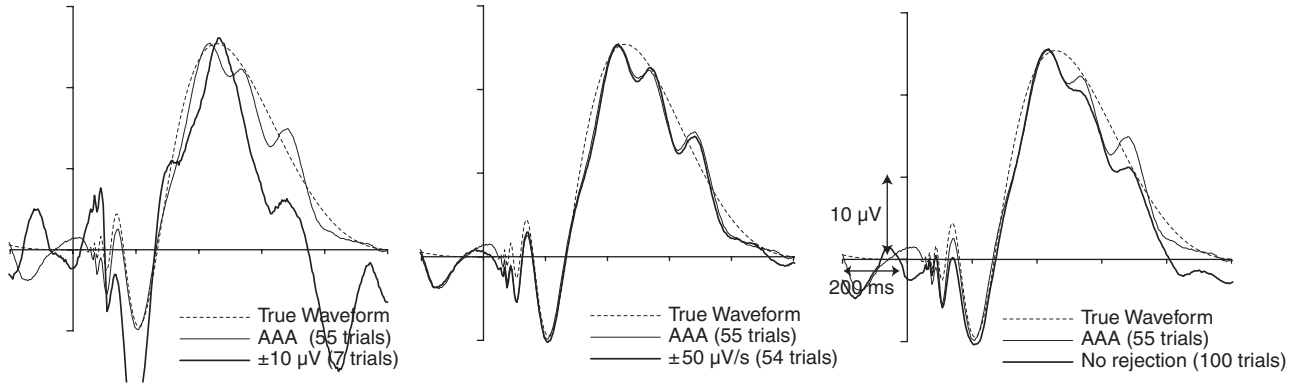

Figure 5. Jump detection simulations. As in Figure 4, these results are obtained in the 32-channel condition, with the true simulated ERP plotted as a reference. The true waveform was estimated most accurately using the single-trial artifact detection method using ± 50 μV, ± 70 μV, and ± 90 μV/s as cutoff values. As shown in the figure, estimation accuracy of the auto-adaptive averaging method was comparable to this result. Using either stricter or more permissive criteria in the single-trial tests resulted in estimates that were in less accurate. AAA: auto-adaptive averaging.

Spikes. Results from the spike-detection simulations are given in Figure 6 and Table 3. In the 32-artifact-channel condition, the auto-adaptive averaging method estimated the true simulated ERP waveform with a slightly higher accuracy than the other methods, with the exception of the no-artifacts test. In all other conditions, performance between the auto-adaptive averaging procedure and the other tests was comparable. Exceptions to this rule can be found in the 4-artifact-channel condition, for the 90-μV single-trial test and 16-artifact-channels condition for the 110-μV and 130-μV single-trials tests, at which performance of the auto-adaptive averaging procedure was somewhat more accurate at estimating the true simulated ERP.

Random walk artifact. Finally, an interesting challenge was posed by the random walk artifact simulating the behavior of an electrode that is about to lose contact. As shown in Figure 7 and in Table 4, the auto-adaptive averaging procedure generally estimated the true simulated ERP waveform with a higher accuracy that of the single-trial artifact detection methods. One exception was formed by the 50-μV single-trial tests, which performed comparably to that of the auto-adaptive averaging procedure in the two-, four-, and eight-artifact channel simulations.

Experimental Data

Figure 8 shows the right-hemisphere-recorded slow-wave potentials that were obtained while participants memorized abstract shapes. These ERPs are reminiscent of the visual working memory-related contralateral slow waves observed previously in the literature (e.g., Drew et al., 2006; Klaver et al., 1999). However, because a number of electrodes were slowly polarizing during the recording session, increasing activity across a relatively wide range of frequencies could be observed. One of the most readily noticeable artifacts was a periodic burst of high-frequency activity, as shown in the first set of averages shown in Figure 8 (top left), which was obtained without artifact rejection. To remove the high-frequency activity, three other ERP averages were made using the moving window peak amplitude detection methods, similar to those used in the simulations described above each using different thresholds. These results were compared to the results provided by the auto-adaptive averaging procedure (bottom).

The application of a 50-μV/4-ms spike detection algorithm changed the ERPs quite drastically, as can be seen in the top and center rows of Figure 8. Increasing the artifact amplitude from 50 μV/4 ms (i.e., detecting potential jumps between two samples) to 500 μV/4 ms resulted in a progressive decrease of the

Table 2. *Comparison of AAA with Single-Trial Artifact Detections in Channel Jump Simulations*

	Number of artifactual channels				
Type	2	4	8	16	32
None	− 4.18	− 3.25	− 4.44	− 10.29	− 17.4
± 10 μV	− 4.78	− 6.86	− 7.44	− 9.33	− 10.4
± 30 μV	− 7.17	− 7.41	− 584	− 6.84	− 9.37
± 50 μV	3.94	8.08	10.1	3.55	2.19
± 70 μV	2.6	5.04	8.69	3.50	3.62
± 90 μV	n.s.	n.s.	8.05	2.61	2.51
± 110 μV	n.s.	n.s.	5.47	n.s.	n.s.
± 130 μV	n.s.	n.s.	2.12	− 2.93	− 4.68
± 150 μV	− 3.07	− 2.59	n.s.	− 4.80	− 10.3

Note. All values are *t* statistics (19 *df*, alpha = .05), comparing the auto-adaptive averaging performance to the single-trial artifact detection methods using the specified cutoff values and number of artifactual channels. Negative *t* values indicate that the auto-adaptive averaging methods estimated the true signal with higher accuracy than the comparison method. AAA: auto-adaptive averaging; None: average obtained without artifact detection, n.s.: not significant.

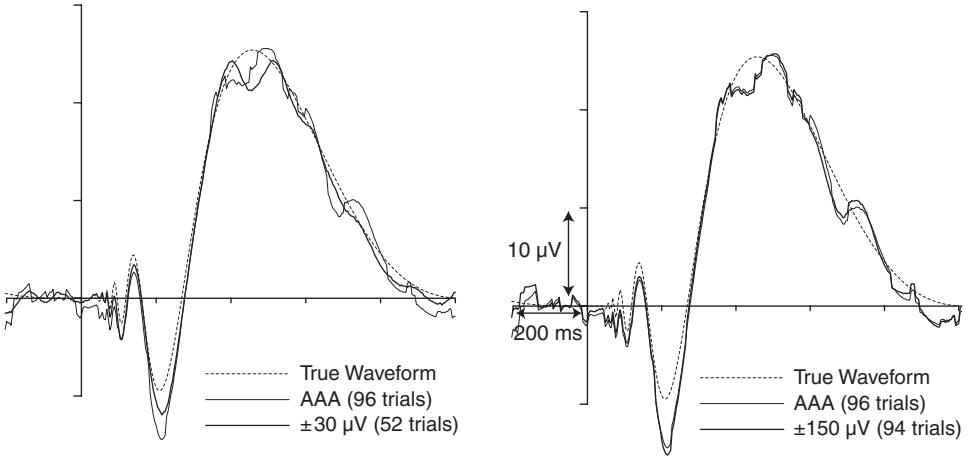

Figure 6. Spike detection simulations. As in previous figures, these results were obtained in the 32-channel simulations. Auto-adaptive averaging performance was somewhat better than the single-trial artifact detection methods. The auto-adaptive averaging method discarded a relatively small number of trials, and, although residual traces of spike activity can be observed in the averaging estimates of both the single $\pm$ 150-μV single-trial detection procedure and the auto-adaptive averaging procedure, the overall distortion on the waveform is still relatively small. AAA: auto-adaptive averaging.

number of trials rejected. Whereas a 50-μV/4-ms criterion resulted in the rejection of a relatively high number of trials, the more liberal rejection criteria eventually resulted in ERPs that resembled the ERPs obtained using the auto-adaptive averaging procedure.

Discussion

This article describes a new method for the detection and rejection of recording artifacts in ERP data. Artifacts are found using an automated procedure that estimates the minimum level of residual noise in the ERP waveform during the average process, and trials having a negative impact on the noise term are excluded. The method was tested by applying it to simulated data, which had one of four different artifact types. In all but one case, the auto-adaptive averaging procedure estimated the simulated ERP signal with a higher than or comparable to accuracy of the most optimally tuned single-trial automated artifact detection methods. Strengths, weaknesses, and possible conditions of the method's use are discussed below.

Performance of the auto-adaptive averaging methods was found to be particularly strong for artifacts that spanned a wide range of frequencies, as demonstrated by the random walk simulations. In addition, performance of the auto-adaptive averaging method was found to be significantly more accurate on the majority of drift detection simulations. This result can be explained to a large degree by the fact that these artifacts exert a relatively long duration influence on the ERP signal, which has the effect that trials containing this type of artifact are easily classified as having a high impact and subsequently as having a negative effect on the residual noise term in the observed ERP waveform.

It is important to stress in this respect that even though comparable performance could be obtained using the single-trial artifact detection methods, this was only the case in a limited number of simulations. In these cases the single-trial artifact rejection thresholds could be considered to be optimally configured. In a fully controlled environment such as the simulations used here, it is possible to compare the ERPs resulting from the various averages to the actual input signal. Hence it is possible to optimize the actual single-trial artifact detection cutoff criteria to the point of near optimal performance. In case of real data, this is oftentimes not the case, due to the fact that the true ERP shape is not a priori known, leading to relatively indirect estimates of the

Table 3. *Comparison of AAA with Single-Trial Artifact Detections in Spike Simulations*

	Number of artifactual channels				
Type	2	4	8	16	32
None	n.s.	n.s.	n.s.	n.s.	n.s.
$\pm$ 10 μV	n.s.	n.s.	n.s.	n.s.	− 2.30
$\pm$ 30 μV	n.s.	n.s.	n.s.	n.s.	− 2.30
$\pm$ 50 μV	n.s.	n.s.	n.s.	n.s.	− 2.30
$\pm$ 70 μV	n.s.	n.s.	n.s.	n.s.	− 2.32
$\pm$ 90 μV	n.s.	− 2.12	n.s.	n.s.	− 2.32
$\pm$ 110 μV	n.s.	n.s.	n.s.	− 2.10	− 2.30
$\pm$ 130 μV	n.s.	n.s.	n.s.	− 2.47	− 2.11
$\pm$ 150 μV	n.s.	n.s.	n.s.	n.s.	− 2.47

Note. All values are *t* statistics (19 *df*, alpha = .05), comparing the auto-adaptive averaging performance to the single-trial artifact detection methods using the specified cutoff values and number of artifactual channels. Negative *t* values indicate that the auto-adaptive averaging methods estimated the true signal with higher accuracy than the comparison method. AAA: auto-adaptive averaging; None: average obtained without artifact detection, n.s.: not significant.

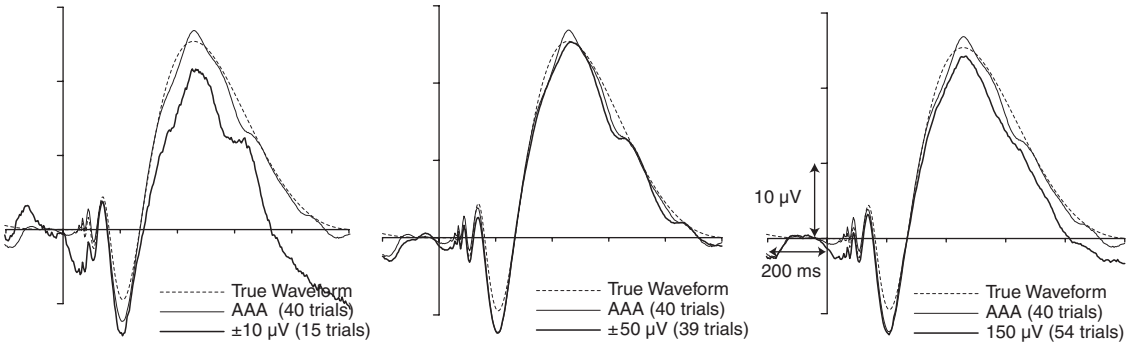

Figure 7. Random walk simulation results. As in previous figures, these results were obtained in the 32-channel simulation. The auto-adaptive averaging procedure was more accurate in estimating the true simulated ERP waveform than the single-trial detection procedures. Somewhat comparable results were obtained by the single-trial detection method using the ± 50-μV threshold. The thresholds resulted in substantial drifts in the estimation of the simulated ERP. AAA: auto-adaptive averaging.

ERPs' signal-to-noise ratio (e.g., Mocks, Gasser, & Tuan, 1984), thus making it impossible to verify the correctness of one's artifact cutoff criteria.

Even in the case where the single-trial methods were configured to near optimal performance, performance of the auto-adaptive averaging method was still somewhat better in cases of long duration artifacts (i.e., in case of the drift and random walk simulations). The working memory data illustrate that the auto-adaptive averaging procedure yielded ERPs that corresponded to the averages that were obtained using the 500-μV and 100-μV/4 ms single-trial cutoff criteria, again suggesting that the optimum configuration would have been somewhere in between these values. Although similar performance could be obtained using the traditional artifact rejection methods, it should be stressed that the latter are based on arbitrary thresholds, which took several passes to configure. In contrast, the auto-adaptive averaging procedure yielded the result plotted in Figure 8 without any user configuration, thereby resulting in unbiased estimates of artifactual activity.

Performance of the auto-adaptive averaging procedure was somewhat less accurate in the detection of spikes and/or other artifacts containing high-amplitude spikes. The reasons for the fact that the traditional single-trial peak-to-peak detection methods outperformed the auto-adaptive averaging procedure are probably threefold. First, due to their transient, high-amplitude characteristics, these artifacts are very easy to detect as long as their amplitude clearly stands out from the background EEG. Second, spikes that have a relatively low amplitude are not likely to have an impact on the ERP that is significant enough to provoke a clear distortion of the waveform. Third, due to their short duration, even residual spike activity that is still present in the ERP will not have sufficient power to affect the multichannel residual power estimates computed by the auto-adaptive averaging procedure.

Three additional cases should be considered where performance of the auto-adaptive averaging procedure could potentially fall short. The first case is where the majority of trials in a particular recording are artifactual, the second case is one where artifacts have a tendency to synchronize with stimulus presentation, and the third case comprises trials in which is EEG signal is dead, that is, when the amplifier has reached the limits of its dynamic range, resulting in a flat line.

Although not included here, some additional simulations were conducted in which large artifacts were present throughout the entire data set. In cases like these, the artifactual activity will become indiscriminable from the background activity, in that artifactual trials will no longer have a significantly larger impact on the ERP signal than any other trial, and therefore it will no longer be possible to determine whether ERPs are artifactual on the basis of their impact. Simulation results also showed that, in this case, performance of the auto-adaptive averaging method was not significantly better than that of single-trial artifact

Table 4. *Comparison of AAA with Single-Trial Artifact Detections in Random Walk Simulations*

	Number of artifactual channels				
Type	2	4	8	16	32
None	− 2.15	− 4.63	− 2.46	− 5.13	− 8.12
± 10 μV	− 5.67	− 7.27	− 5.63	− 5.55	− 6.06
± 30 μV	− 4.13	− 2.93	− 4.10	− 6.54	− 7.14
± 50 μV	n.s.	n.s.	n.s.	− 4.08	− 4.46
± 70 μV	− 2.21	− 3.75	− 3.64	− 6.95	− 4.65
± 90 μV	− 2.20	− 4.95	− 3.24	− 6.12	− 4.65
± 110 μV	− 2.21	− 5.21	− 3.25	− 6.13	− 4.63
± 130 μV	− 2.24	− 5.21	− 3.25	− 6.13	− 4.63
± 150 μV	− 2.24	− 5.21	− 3.25	− 6.12	− 4.63

Note. All values are *t* statistics (19 *df*, alpha = .05), comparing the auto-adaptive averaging performance to the single-trial artifact detection methods using the specified cutoff values and number of artifactual channels. Negative *t* values indicate that the auto-adaptive averaging methods estimated the true signal with higher accuracy than the comparison method. AAA: auto-adaptive averaging; None: average obtained without artifact detection, n.s.: not significant.

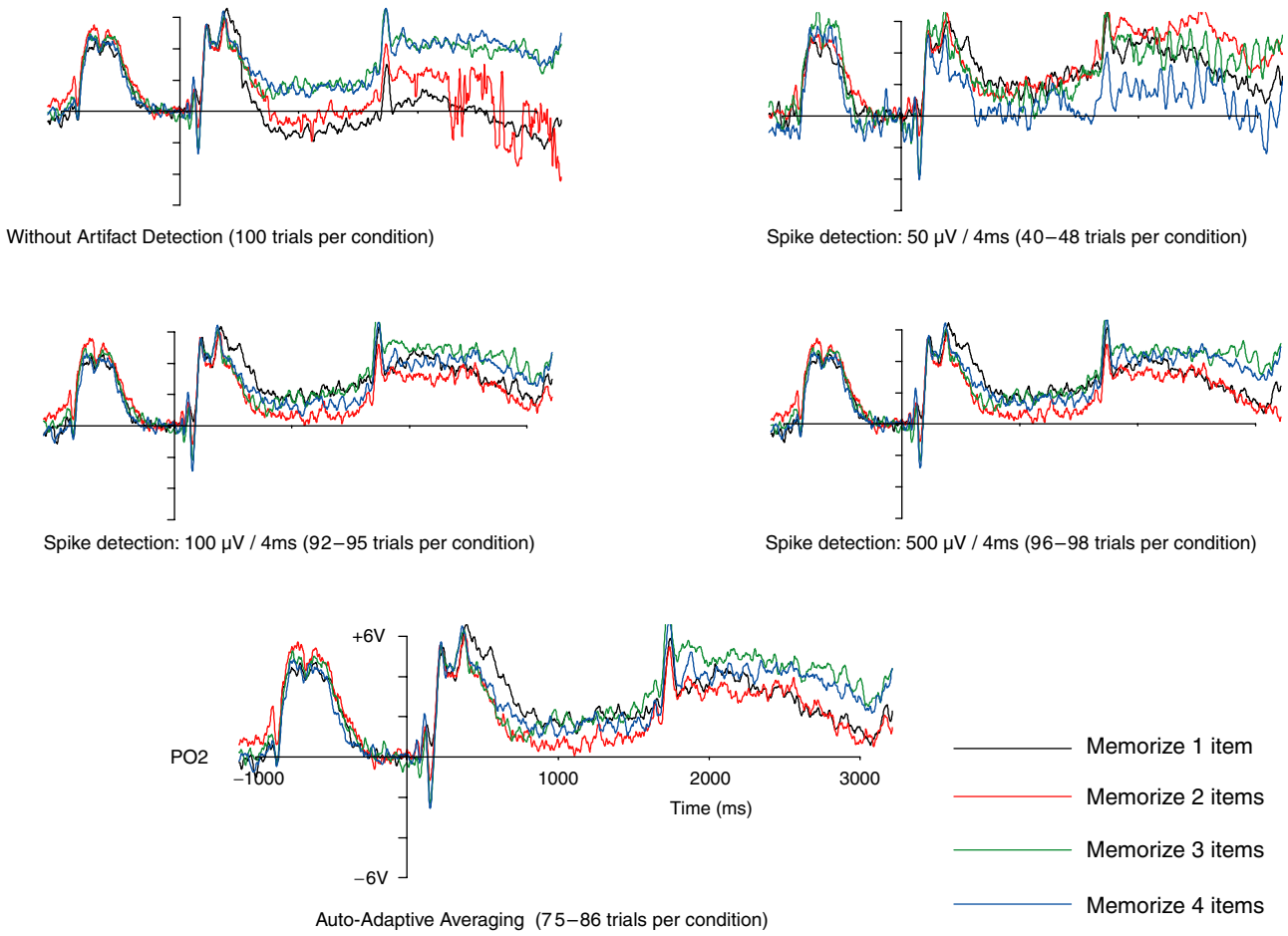

Figure 8. Data from a working memory experiment. Participants were memorizing abstract visual shapes that were presented in the left visual hemifield, and data are shown from a contralateral parieto-occipital recording site (PO_2). High-frequency artifacts were present in the data (top left). After the application of a $\pm$ 50-μV peak-to-peak amplitude test, much of this spike artifact could be removed, but the resulting data were still relatively noisy, due to the high trial-exclusion rate (top right). Using more permissive cutoff criteria eventually resulted in relatively artifact-free ERP data, as shown in the center row. Both averages are comparable to those obtained by the auto-adaptive averaging method, shown at the bottom. Importantly, the results obtained by the auto-adaptive averaging procedure were obtained without the requirement of user-specified thresholds.

detection procedures. It should be cautioned that the use of the auto-adaptive averaging method can, in this respect, *not* serve as a substitute for recording data of sufficient quality.

A second situation that should be considered in this respect is that of an artifact that more or less consistently occurs in synchrony with the event of interest. A good example of such an artifact could be the eyeblink, as some participants have the tendency to synchronize their blinks with the stimulus. On some occasions this can be desirable, as in many cases participants are encouraged to blink sometime after at the end of a trial. In other cases, however, it can be undesirable, in particular when volunteers have problems suppressing reflexive eyeblinks immediately following stimulus presentation. This being the case, the averaging process will not attenuate the artifact by much, and if such artifacts are present on the majority of trials, the impact of artifactual trials will be estimated to be relatively low. For this reason, the temporal characteristics and frequency of occurrence of the artifacts should be monitored. I therefore recommend removal of ocular artifacts from the data, either by rejecting those trials before averaging or by applying one of the linear regression methods (Croft, Chandler, Barry, Cooper, & Clarke, 2005; Gratton et al., 1983; van den Berg-Lenssen et al., 1989; Woestenburg et al., 1983), independent component analysis (Jung et al., 2000), or a multiple source correction technique (Berg & Scherg, 1994).

A third type of artifact that requires special consideration is that of flat lines. Flat lines typically occur after electrode polarization, when the amplifier reaches the limits of its digitization range and becomes saturated. Although trials containing flat lines are clearly artifactual, their impact on the observed ERP signal is relatively low. This seemingly contradictory observation can easily be explained by the fact that the noise component of a singe EEG epoch is considered to be the time-variant part of the background EEG activity, that is, all the activity not related to the ERP. An estimate of the time-variant component can be obtained by subtracting out the observed ERP from the EEG epoch (Gratton et al., 1983; Mocks et al., 1984). In the case of flat-line trials, the estimated noise term would thus equate to the power of the ERP signal itself. Because the estimated power of the ERP is considerably smaller than that of an artifact-free single trial, flat-line trials would be classified as having a relatively low impact.

A final issue that requires discussion concerns the treatment of artifacts across multiple channels. It is common practice in the ERP literature to discard a trial completely even when artifacts occur only on one channel. This practice is important, because differences in the number of trials at each channel could significantly bias the signal-to-noise ratio of ERP waves across channels. These differences could in turn bias scalp topography estimates using spline interpolation (Potts, Dien, Hartry-Speiser, McDougal, & Tucker, 1998) or Laplacian transformations (Tandonnet, Burle, Hasbroucq, & Vidal, 2005). Dipole localization procedures are also particularly sensitive to this type of artifact. For this reason, artifactual trials should always be eliminated in full. In the auto-adaptive averaging method, this is achieved by computing impact and residual noise across multiple channels. This procedure is justified, because in the presence of reasonably clean data, artifacts should have enough power to increase the overall noise estimates, even when even when they occur in just a single channel. The simulation data presented here confirm this assumption, as they indeed show that the auto-adaptive averaging procedure is capable of discarding artifactual trials, even when these artifacts are only present in a subset of data. Again, it should be cautioned that the data should be reasonably clean to begin with. When a small subset of channels is continually sensitive to artifacts, it is advisable to substitute the affected channel by interpolating from a subset of adjacent channels (Picton et al., 2000). In contrast, when problems with such channels occur only intermittently, discarding the affected trials should be sufficient. In particular, when these channels are drifting or displaying random behavior, the auto-adaptive averaging procedure will be able to identify these artifacts, as shown in the simulations and the experimental data.

Summary and Conclusions

This article presents a new method for the automated detection of artifacts in ERP data, based on the analysis of residual noise estimates in the ERP waveform as trials are added to the average. This method, dubbed the auto-adaptive averaging procedure, was found to be able to detect artifacts with accuracy comparable to or higher than single-trial-based artifact detection methods, in particular when the artifact in question has a relatively long duration impact and poorly defined temporal and frequency characteristics. An added advantage is that, provided that the recorded data are reasonably clean, the method is capable of detecting these artifacts without the painstakingly precise tweaking of the single-trial detection parameters that were required to equate the performance of the auto-adaptive averaging procedure.

REFERENCES

beim Graben, P. (2001). Estimating and improving signal-to-noise ratio of time series by symbolic dynamics. *Physical Review E*, *64*, 1–15.

Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology*, *90*, 229–241.

Bosch, V., Mecklinger, A., & Friederici, A. D. (2001). Slow cortical potentials during retention of object, spatial, and verbal information. *Cognitive Brain Research*, *10*, 219–237.

Croft, R. J., & Barry, R. J. (1998). EOG correction: A new perspective. *Electroencephalography and Clinical Neurophysiology*, *107*, 387–394.

Croft, R. J., & Barry, R. J. (2000). EOG correction: Comparing different calibration methods, and determining the number of epochs required in a calibration average. *Clinical Neurophysiology*, *111*, 440–443.

Croft, R. J., & Barry, R. J. (2002). Issues relating to the subtraction phase in EOG artefact correction of the EEG. *International Journal of Psychophysiology*, *44*, 187–195.

Croft, R. J., Chandler, J. S., Barry, R. J., Cooper, N. R., & Clarke, A. R. (2005). EOG correction: A comparison of four methods. *Psychophysiology*, *42*, 16–24.

Drew, T. W., McCollough, A. W., & Vogel, E. K. (2006). Event-related potential measures of visual working memory. *Clinical Electroencephalography and Neuroscience*, *37*, 286–291.

Fisch, B. J. (1991). *Spehlman's EEG primer* (2nd ed). Amsterdam: Elsevier.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484.

Grent-'t-Jong, T., & Woldorff, M. G. (2007). Timing and sequence of brain activity in top-down control of visual-spatial attention. *PLoS Biology*, *5*, 114–126.

Hennighausen, E., Heil, M., & Rosler, F. (1993). A correction method for DC drift artifacts. *Electroencephalography and Clinical Neurophysiology*, *86*, 199–204.

Hickey, C., McDonald, J. J., & Theeuwes, J. (2006). Electrophysiological evidence of the capture of visual attention. *Journal of Cognitive Neuroscience*, *18*, 604–613.

Hopf, J. M., & Mangun, G. R. (2000). Shifting visual attention in space: An electrophysiological analysis using high spatial resolution mapping. *Clinical Neurophysiology*, *111*, 1241–1257.

Johnson, R., Jr. (1989). Developmental evidence for modality-dependent P300 generators: A normative study. *Psychophysiology*, *26*, 651–667.

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, *111*, 1745–1758.

Klaver, P., Talsma, D., Wijers, A. A., Heinze, H. J., & Mulder, G. (1999). An event-related brain potential correlate of visual short-term memory. *NeuroReport*, *10*, 2001–2005.

Mocks, J., Gasser, T., & Tuan, P. D. (1984). Variability of single visual evoked potentials evaluated by two new statistical tests. *Electroencephalography and Clinical Neurophysiology*, *57*, 571–580.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2002). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cognitive Brain Research*, *15*, 115–128.

Niedermeyer, E., & Lopez-De Silva, F. (1993). *Electroencephalography. Basic principles, clinical application, and related fields* (3rd ed.). Baltimore, MD: Williams and Wilkins.

Papoulis, A. (1991). *Probability, random variables, and stochastic processes* (3rd ed.). New York: McGraw-Hill.

Pelosi, L., & Blumhardt, L. D. (1999). Effects of age on working memory: An event-related potential study. *Cognitive Brain Research*, *7*, 321–334.

Perlin, K. (2002). Improving noise. *Computer Graphics*, *35*.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., et al. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*, 127–152.

Potts, G. F., Dien, J., Hartry-Speiser, A. L., McDougal, L. M., & Tucker, D. M. (1998). Dense sensor array topography of the event-related potential to task-relevant auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, *106*, 444–456.

Ruchkin, D. S., Berndt, R. S., Johnson, R., Ritter, W., Grafman, J., & Canoune, H. L. (1997). Modality-specific processing streams in verbal working memory: Evidence from spatio-temporal patterns of brain activity. *Cognitive Brain Research*, *6*, 95–113.

Scheffers, M. K., Johnson, R., Jr., & Ruchkin, D. S. (1991). P300 in patients with unilateral temporal lobectomies: The effects of reduced stimulus quality. *Psychophysiology*, *28*, 274–284.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*, 679–690.

Talsma, D., Kok, A., & Ridderinkhof, K. R. (2006). Selective attention to spatial and non-spatial visual stimuli is affected differentially by age: Effects on event-related brain potentials and performance data. *International Journal of Psychophysiology*, *62*, 249–261.

Talsma, D., Wijers, A. A., Klaver, P., & Mulder, G. (2001). Working memory processes show different degrees of lateralization: Evidence from event-related potentials. *Psychophysiology*, *38*, 425–439.

Talsma, D., & Woldorff, M. G. (2005a). Methods for the Estimation and removal of artifacts and overlap in ERP waveforms. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 115–148). Cambridge, MA: MIT Press.

Talsma, D., & Woldorff, M. G. (2005b). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, *17*, 1098–1114.

Tandonnet, C., Burle, B., Hasbroucq, T., & Vidal, F. (2005). Spatial enhancement of EEG traces by surface Laplacian estimation: Comparison between local and global methods. *Clinical Neurophysiology*, *116*, 18–24.

Van den Berg-Lenssen, M. M. C., Brunia, C. H. M., & Blom, J. A. (1989). Correction of ocular artifacts in EEGs using an autoregressive model to describe the EEG—A pilot study. *Electroencephalography and Clinical Neurophysiology*, *73*, 72–83.

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, C. V. (1964). Contingent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. *Nature*, *203*, 380–384.

Woestenburg, J. C., Verbaten, M. N., & Slangen, J. L. (1983). The removal of the eye-movement artifact from the EEG by regression analysis in the frequency domain. *Biological Psychology*, *16*, 127–147.

Yago, E., Escera, C., Alho, K., & Giard, M. H. (2001). Cerebral mechanisms underlying orienting of attention towards auditory frequency changes. *NeuroReport*, *12*, 2583–2587.

(Received March 26, 2007; Accepted May 14, 2007)

APPENDIX: SIGNAL-TO-NOISE CALCULATIONS

In signal analysis, it is considered that a measured time series $x(t)$ consists of a deterministic signal $s(t)$, and some additional noise $\varepsilon_\sigma(t)$ with variance σ^2:

$$x(t) = s(t) + \varepsilon_\sigma(t). \tag{1}$$

In event-related potential analysis, $s(t)$ is considered to be the time-locked brain response, which is embedded in the brain's spontaneous background activity and recording artifacts, both being described by $\varepsilon_\sigma(t)$. The time-locked part, $s(t)$, is obtained by averaging across an ensemble of EEG epochs $x_i(t)$, where i represents the ensemble index ranking across all trials N, $1 \leq i \leq N$:

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t). \tag{2}$$

A key objective of ERP signal averaging is to obtain an optimally clean signal, that is, to obtain the highest possible signal-to-noise ratio. The signal-to-noise ratio is given by the ratio of signal power over noise power (Papoulis, 1991):

$$Q = \sqrt{\frac{P_s}{P_n}}. \tag{3}$$

In event-related potential research it is typically impossible to directly observe P_S and P_N, as a single-trial waveform consists of both the signal and the noise (with the latter being composed of both background EEG, equipment noise, and recording artifacts). A commonly used statistical estimate of P_S and P_N was developed by Mocks, Gasser, and Tuan (1984), according to the following equations:

$$\hat{P}_{\mathrm{s}} = \frac{1}{T}\int_0^T \bar{x}^2(t)\,dt - \frac{1}{N}\hat{P}_N \tag{4}$$

$$\hat{P}_N = \frac{1}{N-1}\sum_{i=1}^{N}\frac{1}{T}\int_0^T (x_i(t) - \bar{x}(t))^2. \tag{5}$$

In Mocks' estimate, signal power P_S is considered to be the power of the ERP waveform after subtracting out an estimated noise contribution, with the latter consisting of $1/N$ times the estimate of the average noise level across all EEG epochs N, $1 \leq i \leq N$. The noise term itself is therefore considered to be the average power of the individual EEG epochs, after subtracting out the event-related potential on each trial.

If the noise term is constant across trials and is neither correlated with the signal nor with itself across trials, it follows from Equations (4) and (5) that averaging N trials yields a signal-to-noise ratio improvement of $\sqrt{N}$. Because the ERP signal is assumed to remain constant across trials, this improvement of the signal-to-noise ratio can be equated to a $1/\sqrt{N}$ reduction in estimated noise in the obtained ERP averages.

Figure 2a shows this theoretically expected $\sqrt{N}$ noise decrease of the ERP signal as a function of the numbers of trials included in the average. In the presence of artifactual trials, it can no longer be expected that the noise levels in the ERP signal decrease according to the $1/\sqrt{N}$ rule, due to the fact that the noise levels are no longer equal across trials. It follows from Equation (5) that P_n is obtained by averaging the integrated power of the individual trials after subtracting of the (event-related) average signal. Therefore, under the assumption that the ERP signal remains constant, an estimate of the noise reduction $NR_{(Ns)}$ across a subset of N, Ns, $1 \leq i \leq Ns$ epochs can be obtained using the following equation:

$$NR_{(Ns)} = \frac{\frac{1}{Ns}\sum_{i=1}^{Ns}\hat{P}_N(i)}{\sqrt{Ns}} \tag{6}$$

where Ns denotes the size of the subset of N, $\hat{P}_N(i)$ the estimated power of the EEG background noise on trial i, and $NR_{(Ns)}$ the noise estimate obtained by averaging the first N trials. Because the ERP is assumed to be time invariant across trials, $\hat{P}_N(i)$ can be estimated by applying Equation (5) across the subset of trials Ns, but using an estimate $\bar{x}(t)$ that is based on the full set of trials N.